Applied Statistics:

9/3/20

Books:

1) Johnson, Beaker book.
2) ISLR, Springer.

Data frames:

10/3/20

$$\begin{cases} n : \text{statistical units.} \\ p : \text{features.} \end{cases}$$

$$\underline{x}_1 = (x_{11}, x_{12}, \dots, x_{1p})' \in \mathbb{R}^p$$

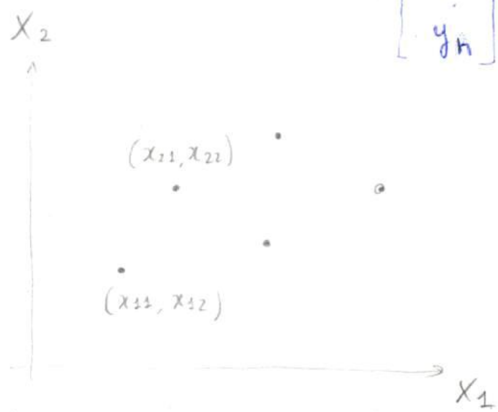
$$\underline{x}_2 = (x_{21}, \dots, x_{2p})'$$

$$\vdots$$

$$\underline{x}_n = (x_{n1}, \dots, x_{np})'$$

Basically, this is a matrix of observations:

$$\text{stat. units:} \quad \begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ n \end{array} \begin{array}{c} X_1 \quad\quad X_2 \quad\quad\quad\quad X_p \to \text{features} \\ \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & \dots & \dots & x_{2p} \\ \vdots & & & \vdots \\ x_{n1} & \dots & \dots & x_{np} \end{bmatrix} \end{array} = \mathbb{X} \quad \underline{\text{data matrix}} \ (\text{data frame}) \qquad \begin{array}{c} Y \\ \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \end{array}$$



In our usual setting, $\boxed{n >> p}$

$\underline{X}$: random vector, $\underline{X} \in \mathbb{R}^p$.

Y : random variable, $Y \in \mathbb{R}$.

Problem: <u>use $\underline{X}$ to predict Y.</u>

We are looking for the "<u>best</u>" function $f : \mathbb{R}^p \to \mathbb{R}$ to predict Y in terms of $\underline{X}$.

Best in the sense that $\boxed{\mathbb{E}\left[\left(Y - f(\underline{X})\right)^2\right]}$ should be <u>minimum</u>.

We are taking the squares because we want to minimize the <u>mean squared error</u>. We could use different measures, like the absolute value. For now, we are working in L^2.

$E[(Y-f(\underline{X}))^2]$.

Exercise: find (prove) that: $\boxed{\underset{k}{\mathrm{Argmin}}\ E[(Y-k)^2] = E[Y].}$

In order to solve the more general problem of finding f, we do:

$$E[(Y-f(\underline{X}))^2] = E[(Y - E[Y|\underline{X}] + E[Y|\underline{X}] - f(\underline{X}))^2] =$$

$$= E[(Y-E[Y|\underline{X}])^2] + E[(E[Y|\underline{X}] - f(\underline{X}))^2] +$$

$$+ \underbrace{2E[(Y-E[Y|\underline{X}])(E[Y|\underline{X}] - f(\underline{X}))]}_{(*)}.$$

We prove that: given W, Z,

$$E[W] = E[E[W|Z]].$$

Hence the middle term becomes:

$$(*) \underset{\text{condition on } \underline{X}}{=} E[(E[Y|\underline{X}] - f(\underline{X}))\underbrace{E[Y - E[Y|\underline{X}]\,|\,\underline{X}]}_{= E[Y|\underline{X}] - E[Y|\underline{X}] = 0}] \Rightarrow (*) = 0.$$

Then:

$$\boxed{E[(Y-f(\underline{X}))^2] = E[(Y-E[Y|\underline{X}])^2] + E[(E[Y|\underline{X}] - f(\underline{X}))^2].}$$

Thus, our best guess is to take $\boxed{f(\underline{X}) = E[Y|\underline{X}].}$ solution of the optimization problem

The first term <u>cannot be eliminated</u>; it's like a constant.
There will always be a difference between Y and its prediction $f(\underline{X})$:

$$\boxed{Y - f(\underline{X}) = \varepsilon}$$

This relation is giving us a <u>model</u>:

$$Y = f(\underline{X}) + \varepsilon, \quad \text{where} \quad f(\underline{X}) = E[Y|\underline{X}].$$

Which are the features of ε?

Observe that:

$$E[Y] = E[E[Y|\underline{X}]] + E[\varepsilon]$$
$$\| $$
$$= E[Y] + E[\varepsilon] \Rightarrow \boxed{E[\varepsilon] = 0.}$$

We want to use data to estimate f.

Say that $\hat{f}$ is my estimate of f. I want to see how good is this estimate:

$$\begin{cases} \hat{f} : \text{estimate of } f \text{ (via } \mathbb{X}). \\ \underline{x}_0 \in \mathbb{R}^p \xrightarrow[\hat{f}]{\text{prediction}} Y_0 . \end{cases} \qquad Y_0 = f(\underline{x}_0) + \varepsilon_0 .$$

I want to find $\underbrace{E|_{\mathbb{X}}}_{\text{conditioning on data}} \left[(Y_0 - \hat{f}(\underline{x}_0))^2 \right] =$

$$= E|_{\mathbb{X}} \left[(f(\underline{x}_0) + \varepsilon_0 - \hat{f}(\underline{x}_0))^2 \right] =$$

$$= E|_{\mathbb{X}} \left[\underbrace{(f(\underline{x}_0) - \hat{f}(\underline{x}_0))^2}_{\text{constant}} \right] + \underbrace{E|_{\mathbb{X}} \left[\varepsilon_0^2 \right]}_{\varepsilon_0 \text{ independent on } \mathbb{X}} + 2 E|_{\mathbb{X}} \left[\underbrace{(f(\underline{x}_0) - \hat{f}(\underline{x}_0))}_{\text{constant, can take out of the expectation}} \varepsilon_0 \right]$$

$$= (f(\underline{x}_0) - \hat{f}(\underline{x}_0))^2 + Var(\varepsilon_0) + 0 .$$

$$\Rightarrow \boxed{E|_{\mathbb{X}} \left[(Y_0 - \hat{f}(\underline{x}_0))^2 \right] = \underbrace{(f(\underline{x}_0) - \hat{f}(\underline{x}_0))^2}_{\text{reducible term}} + \underbrace{Var(\varepsilon_0)}_{\text{irreducible error term}} .}$$

reducible term: if $\hat{f}$ is a good estimate, it will be small.

irreducible error term

Estimation of f:

$\hat{f}(\underline{x})$ estimate of $E[Y | \underline{X}]$.

One option is the following:

$\hat{f}(x)$: average of $\{y_i\}$ where $x_i \in \mathcal{N}(\bar{x})$.

When p is large, this technique is not any longer feasible $\Rightarrow$ "Curse of dimensionality".

Curse of dimensionality:

$p = 1$:

$x \sim \mathcal{U}[S^1(1)]$.

$S^1(1) \rightsquigarrow$ sphere of radius 1 in dimension 1

How big is the distance that you have to travel in order to capture 10 % of your friends x?

$$0.1 = \frac{\text{length}(S^1(r))}{\text{length}(S^1(1))} = \frac{2r}{2} = r .$$